Thailand Statistician
July 2006; 4: 27-41
http://statassoc.or.th
Contributed paper

A Bayesian Approach for Fitting a Random Effect Differential Item Functioning Across Group Units

Saengla Chaimongkol*[a], Fred W. Huffer [b], and Akihito Kamata [c]
[a] Department of Mathematics and Statistics, Thammasat University, Phathum Thani, 12121, Thailand.
[b] Department of Statistics, College of Arts & Sciences, Florida State University, Florida, U.S.A.
[c] Department of Educational Psychology & Learning Systems, College of Education, Florida State University, Florida, U.S.A.
***Author for correspondence**, e-mail: saengla@mathstat.sci.tu.ac.th

Received: 6 April 2006
Accepted: 18 July 2006

Abstract

This study proposed a multilevel logistic regression model to evaluate variation of differential item functioning (DIF). The model accounts for the three level nested structure of the data and combines results of logistic regression analyses to investigate the variation of DIF across level-3 units. A simulation study is presented to assess the adequacy of the proposed model. The parameters of the proposed model were estimated by using a Bayesian approach implemented by the WinBUGS 1.4.

Keywords: Bayesian, level3-units, multilevel logistic regression, random effect DIF, WinBUGS.

1. Introduction

DIF is presented for a test item when respondents from two subpopulations with the same trait level have different probability of answering the item correctly. A consequence of having a DIF item is that the same true trait levels for examinees from different subpopulations could indicate different total test scores or trait level estimates. Currently, many statistical techniques have been proposed, based upon various theoretical backgrounds and practical purposes. A thorough review of DIF detection methods can be found in Millsap and Everson [6].

Once an item is identified as functioning differently from one subpopulation to another, understanding why the item is functioning differently between groups may be useful for many audiences. As one attempt, Gierl et al. [4] studies gender DIF in mathematics by combing substantive and statistical analyses, as a two-stage process. Three difference statistical methods: SIBTEST, DIMTEST, and multiple linear regression, were used to test hypotheses about gender differences and to test whether content and cognitive differences were among items. Bolt [2], for another example, found that multiple-choice items had more DIF characteristic than constructive-response items between males and females on SAT math pretest items. These results can possibly provide suggestions that may be informative to minimize DIF items in future by many different means, including instruction, policy and test construction. These studies were based on multidimensional IRT based approaches.

As another statistical approach, Swanson et al. [8] proposed a two-level logistic regression model to evaluate sources of DIF. This approach explicitly accounts for the nested structure of the data and combines results of logistic regression analyses across

---

This paper was presented at The Conference of Statistics and Applied Statistics; 2006, 25-26 May 2006, organized by Department of Mathematics and Statistics, Faculty of Science and Technology, Thammasat University, Pathum Thani, 12121, Thailand.

individual items to investigate the variation of DIF. Their level-1 model is a logistic regression model for DIF detection proposed by Swanminathan and Rogers [7]. In the level-2 models, the coefficients from level-1 model are treated as random variables and allow one to incorporate item characteristic variables to the models in order to explain the variation of DIF across items.

There is also a possibility that the magnitude of the DIF varies across group units, such as schools, and communities. Kamata and Binici [5] first attempted to extend a two-level DIF model to three-level model using the hierarchical generalized linear model (HGLM) framework. Their three-level model approach can be used to model variation of DIF across school as well as applied to identify the school characteristic variables that explain such variation. Their models were implemented by the HLM-5 software, which uses the penalized or predictive quasi-likelihood (PQL) method. They found that the variance estimates produced by the HLM-5 for the level 3 parameters are substantially negatively biased. This study extends their work by using a Bayesian approach to obtain more accurate parameter estimates. More specifically, this study will demonstrate a model in such a way that DIF of a particular item may vary among level-3 units.

2.1 Model Specification

To set the notation, let i denote the level-3 units (schools), j denote the level-2 units (students), and k denote the level-1 units (items). Assume that $i = 1,2,...,s$, $j = 1,2,...,n_i$ and $k = 1,2,...,t$. Let y_{ijk} be the dichotomous response with code 1 if student j in school i responds to item k correctly and 0 otherwise. For all i, j and k , y_{ijk} are assumed to be independent Bernoulli random variables with the probability of correct response $p_{ijk} = P(y_{ijk} = 1)$. The random effect DIF model can be written as:

$$y_{ijk} \sim \text{Bernoulli}(p_{ijk})$$

$$\text{logit}(p_{ijk}) = u3_i + u2_{ij} + \alpha 0 G_{ij} - \beta_k - \gamma 0_k G_{ij} + u4_{ik} G_{ij}, \quad (1)$$

where

- $u3_i$ is the random effect for school i. It is assumed to be normally distributed with zero mean and constant variance (i.e., $u3_i \sim N(0,(\sigma 3)^2)$

- $u2_{ij}$ is the ability of student j in the school i. It is a random effect and is assumed to be normally distributed with a non-zero mean and a group-specific variance (i.e., $u2_{ij} \sim N(\mu, \sigma_G^2)$).

- $\alpha 0$ is the fixed effect of belonging to the focal group compared to the reference group, i.e., it is the mean difference between the focal and reference group.

- G_{ij} is the group indicator that either indicates the reference group or the focal group. G_{ij} equals 0 if student j in school i belongs to the reference group and equals 1 if student j in school i belongs to the focal group.

- β_k is a fixed effect representing the difficulty of item k for the reference group.

- $\gamma 0_k$ represents the overall mean DIF for item k across schools

- $u4_{ik}$ is a random increment to the DIF for item k in school i. It is assumed to be normally distributed with mean zero and item-specific variance (i.e., $u4_{ik} \sim N(0,(\sigma 4_k)^2)$). It is further assumed that the random effects $u3_i$, $u2_{ij}$, and $u4_{ik}$ are assumed to be mutually independent.

The values $\sigma 4_k$ provide a set of indices that describe how the DIF varies across schools. A large value of $\sigma 4_k$ indicates that, after controlling for school and student abilities, the DIF varies a great deal from school to school. On the other hand, a small or zero value of $\sigma 4_k$ indicates that the DIF varies little from school to school.

The model (1) is not identified because a constant can be added to all $u2_{ij}$ and all β_k, but the logit of the model does not change. Similarly, a constant can be added to all $\alpha 0$ and all of the $\gamma 0_k$ without changing the logits. To solve the non-identification problem, Bafumi et al. [1] suggested replacing the model parameters with new (adjusted) quantities that are well-identified but preserve the logit of the model. This study adopts

Bafumi et al.'s general approach to identify the model (1) by defining the model parameters as:

$$\begin{aligned} u3_i^{adj} &= u3_i - \bar{u}3 \\ u2_{ij}^{adj} &= u2_{ij} - \bar{\beta} + \bar{u}3 \\ \alpha 0^{adj} &= \alpha 0 - \bar{\gamma} 0 \\ \beta_k^{adj} &= \beta_k - \bar{\beta} \\ \gamma 0_k^{adj} &= \gamma 0_k - \bar{\gamma} 0. \end{aligned} \tag{2}$$

These adjusted quantities will be used in place of the original quantities. The random effect DIF model now can be defined as:

$$\text{logit}(p_{ijk}) = u3_i^{adj} + u2_{ij}^{adj} + \alpha 0^{adj} G_{ij} - \beta_k^{adj} - \gamma 0_k^{adj}{}_k G_{ij} + u4_{ik} G_{ij} \tag{3}$$

2.2 Estimation with WinBUGS

The random effect DIF model can be easily implemented in WinBUGS, an available software for Bayesian analysis, using Gibbs sampling. The parameters in the model are fixed effects ($\alpha 0$, β_k , and $\gamma 0_k$), μ , the random effects ($u3, u2, u4$), and the standard deviation parameters ($\sigma 2_G, \sigma 3,$ and $\sigma 4_k$). The parameters of the greatest interest for the random effect DIF model are the standard deviation of the random DIF magnitude, $\sigma 4_k$. If the $\sigma 4_k$ is large, it indicates that the DIF magnitude of item k is different across schools.

The Bayesian approach treats all unknown parameters as random quantities with appropriate prior distributions. Estimation is based on the joint posterior distribution $P(\theta \mid y)$ where θ is the vector of unknown model parameters (i.e., $\theta = (\{\alpha 0\}, \{\beta_k\}, \{\gamma 0_k\}, \mu, \{\sigma 2_G\}, \sigma 3, \sigma 4_k)$) and y is the sample data. The posterior distribution of θ is obtained from Bayes' theorem as:

$$P(\theta \mid y) = \frac{P(y \mid \theta)P(\theta)}{\int P(\theta)P(y \mid \theta)d\theta}$$
$$\propto P(y \mid \theta)P(\theta).$$

where $P(y \mid \theta)$ is the likelihood and $P(\theta)$ is the prior.

The posterior distribution for θ is proportional to the likelihood multiplied by the prior distribution. Since the item responses given the school and student ability are assumed to be independent, the likelihood for the random effect DIF model is given as:

$$P(y \mid \theta) = \int g(u4;0,\sigma 4_k)\int g(u3;0,\sigma 3)\int g(u2;\mu,\sigma 2_G)\prod_{i,j,k} f(y_{ijk} \mid u2_{ij}, u3_i, u4_{ik})du2_{ij}\, du3_i\, du4_{ik},$$

where

$$f(y_{ijk} \mid u2_{ij}, u3_i, u4_{ik}) = \left(\frac{1}{1+e^{-\eta_{ijk}}}\right)^{y_{ijk}} \left(\frac{e^{-\eta_{ijk}}}{1+e^{\eta_{ijk}}}\right)^{1-y_{ijk}},$$

and $g(u4;0,\sigma 4_k)$, $g(u3;0,\sigma 3)$, and $g(u2;\mu,\sigma 2_G)$ are multivariate normal density of $u4 = \{u4_{ik}\}$, $u3 = \{u3_i\}$, and $u2 = \{u2_{ij}\}$, respectively, $\eta_{ijk} = \text{logit}(p_{ijk})$ from equation (1).

2.3 Choice of Prior Distributions and Specification of Initial Values

Bayesian estimation of the model parameters requires the specification of a prior distribution for all the unknown parameters. In our prior distribution, we use a noninformative, but proper prior distribution. We assume the fixed effect ($\{\alpha 0\}, \{\beta_k\}, \{\gamma 0_k\}$), and μ of $u2$ are independent and normally distributed with mean zero and a huge variance 10^4 $(N(0,10^4))$. For the variance parameters, we follow the recommendation of Gelman [3] that suggests the use of a noninformative uniform prior density on standard deviation parameters unless a weakly informative prior is desired, in which case a half-t family such as a half-Cauchy prior distribution on σ is recommended instead of a uniform prior. The half-t distribution can be defined as the

ratio of the absolute value of a normal random variable centered at 0 and the square root of a gamma random variable. Further details on this distribution can be seen in Gelman [3]. For $\sigma 2_G$ and $\sigma 3$ a noninformative proper uniform prior density with a wide range (i.e., unif(0, 1000)) are used. For the between-school standard deviation of DIF parameter for each item ($\sigma 4_k$), half-Cauchy prior distributions with scale parameter (ξ) of 25 as recommended by Gelman [3] are used.

After setting the prior density for all unknown parameters, the model is completely specified in WinBUGS. Then WinBUGS loads the data and compiles the model. When the model compiles successfully, WinBUGS will load the initial values in the next step. For the random DIF model, we specify 0 as initial value for the fixed effects ($\{\alpha 0\},\{\beta_k\},\{\gamma 0_k\}$) and μ of $u2$, and 1 as the initial value for the standard deviation parameters of ($\sigma 2_G, \sigma 3, \sigma 4_k$). Using Gelman's [3] approach to implementing the half-cauchy prior, the values of $\sigma 4_k$ are actually represented as

$$\sigma 4_k = \frac{|\xi_k|}{\sqrt{\tau_k}}$$

where $\xi_k \sim N(0,25)$ and $\tau_k \sim \chi_1^2$. We assign ξ_k and τ_k initial values of 1, so that $\sigma 4_k$ is also initially equal to 1. The initial values for the random effects $(u2, u3, u4)$ are generated by WinBUGS itself.

When the initial values have been loaded or generated by WinBUGS successfully, WinBUGS now is ready to run the Gibbs sampling to obtain statistical inferences for the unknown parameters. In each situation we study, only one chain is run and the chain is run for 11,000 iterations with a burn-in of 1,000.

3. Simulation Study and Results

3.1 Simulation Design

The simulation study consists of three conditions that vary the number of students per school (n) and number of schools (s): (a) n=50, s=20, (b) n=20, s=50, and (c) n=40, s=40. For each of the three conditions, we simulate 100 data sets from the model (1). For all these simulations, the number of test items is fixed at 10 with the difficulties β_k = -1.0, -1.0, 0, 0, 0, 0, 0, 0, 1.0, 1.0. The mean difference between the reference and focal group is $\alpha 0$ = -1.0. Three situations for DIF will be considered. First, DIF exists but is not consistent across schools. Second, DIF exists and is consistent across schools. Third, DIF is negligible overall but varies from school to school. In order to study these three situations, for items 3, 4, and 5, we take $\gamma 0_k$ to be 0.7, 0.7, and 0.4, respectively, and $\sigma 4_k$ to be 1.0, 0.2, and 1.0, respectively. Then items 3, 4, and 5 represent the first, second, and third situation for DIF, respectively. For the others items, we take $\gamma 0_k$ = 0 and $\sigma 4_k$ = 0.2. The ability of students ($u2_{ij}$) is sampled from N(0,1). The ability of schools ($u3_i$) is sampled from N(0,4). The random effects $u2_{ij}$, $u3_i$, and $u4_{ik}$ are generated independently in each data set. It is assumed that $u2_{ij}$, $u3_i$, and $u4_{ik}$ are mutually independent. Splus is used to create the simulated data, and WinBUGS is used for the subsequent analysis.

3.2 Simulation Results

For each data set and analysis, our posterior inference is based on the output of a Gibbs sampler. We illustrate some typical Gibbs sample output using one data set from the first condition (n=50, s=20). Sample history plot (trace plot), autocorrelation plot and posterior density plot are given for selected parameters for 10,000 iterations after eliminating the first 1,000 iterations. These plots are shown in Figures 1 to 3, respectively.

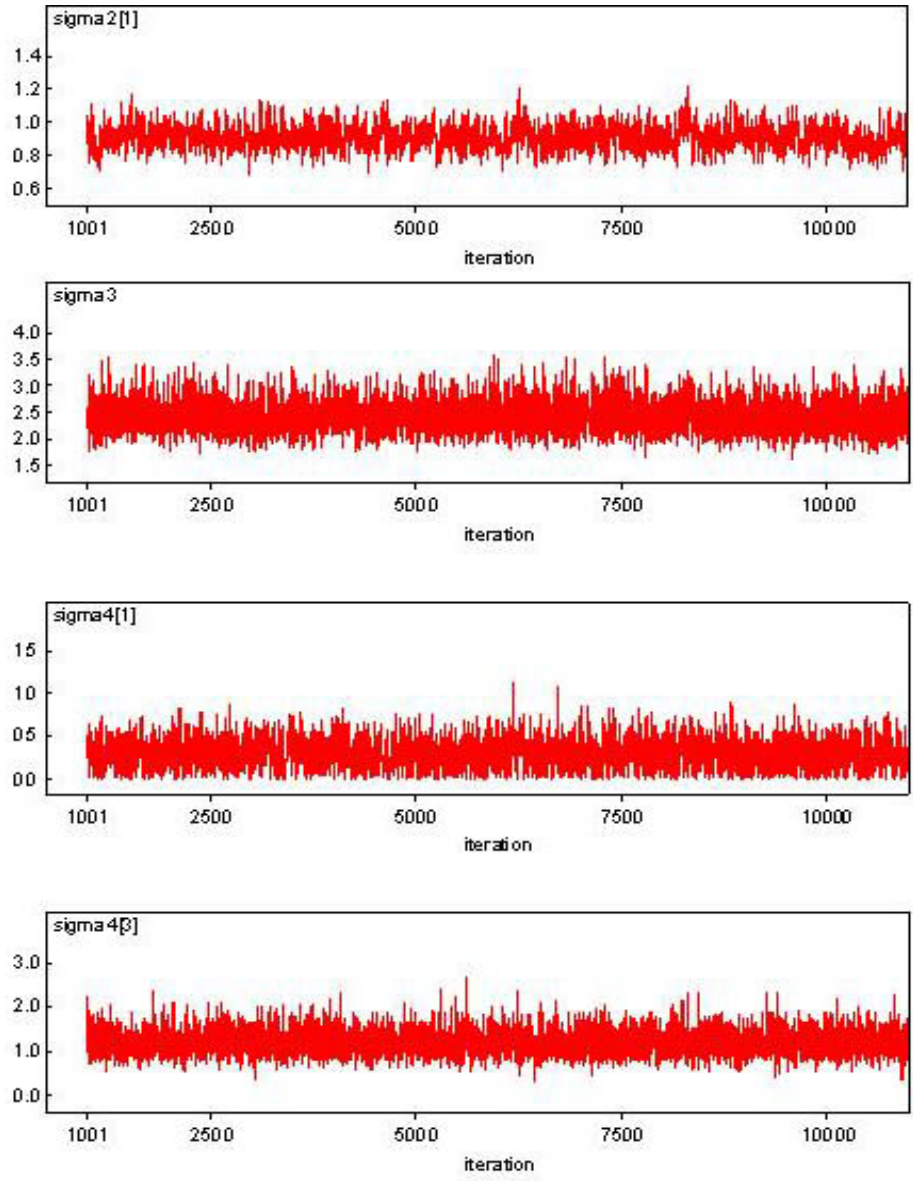

Figure 1. Gibbs sampling trace plots for some standard deviation parameters under condition 1 ($n = 50,\ s = 20$).

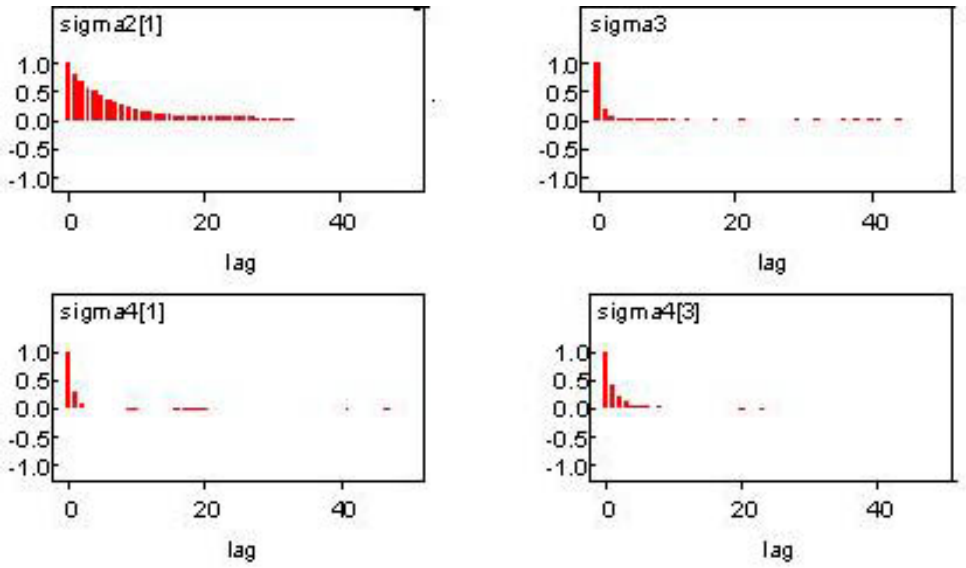

Figure 2. Gibbs sampling autocorrelation plots for some standard deviation parameters under condition 1 ($n = 50,\ s = 20$).

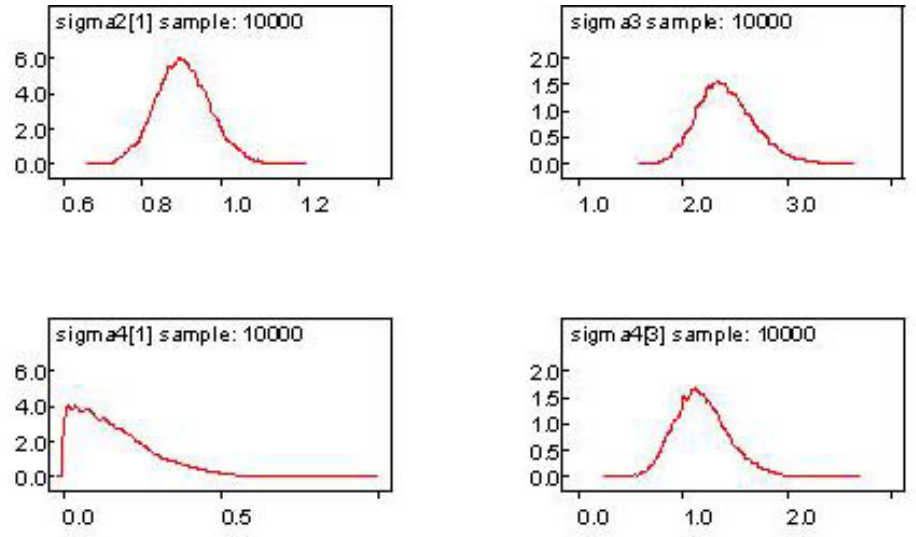

Figure 3. Gibbs sampling density plots for some standard deviation parameters under condition 1 ($n = 50,\ s = 20$).

The trace plots are shown in Figure 1. Each parameter of interest becomes stationary by 1,000 iterations, indicating that convergence has been reached by 1,000 iterations. The autocorrelation plots (Figure 2) show that for all parameters, except the level-2 standard deviations, the autocorrelations decrease to near zero in fewer than l0 lags. The autocorrelations of the level-2 standard deviations approach to near zero by about lag 20. This indicates that the correlation between any two values separated by 10 or more iterations is close to zero, and these values can be treated as being roughly independent. These autocorrelation plots suggest that the chains are mixing well and quickly. In other words, the chains rapidly explore the entire posterior distribution.

The density plots for parameters (Figure 3) show unimodal distributions which are nearly symmetric, and look close to normal except for the plot of $\sigma 4_1$, which has a long right tail and a high peak close to zero. This result is likely due to the values of $\sigma 4_1$ being close to zero, the lower boundary of the parameter space.

Each Gibbs sampler run produces 10,000 values for each parameter in the model. The sample mean and standard deviation (SD) of these 10,000 values estimate the posterior mean and standard deviation for that parameter. For each of the three conditions, the estimates of the posterior mean, standard deviation were computed for all 100 simulated data sets. These results are summarized in Table 1. From this table, we see that the mean over the 100 data sets of standard deviation parameters is far from the true parameter value used in the simulation. Substantial relative bias exists when $\sigma 4_k$ is small (0.2 in our simulations). In order to study the effect of varying the number of schools and the number of students per school on the bias, the relative bias of σ is computed using $(\bar{\hat{\sigma}} - \sigma)/\sigma$. The estimates of the relative bias for all the standard deviation parameters under each condition are presented in Table 2. It can be seen immediately from this table that the point estimates of the standard deviations are positively biased, except for the $\sigma 2_2$ (This exception are likely due to random variation.)

The standard deviations of level-2 ($\sigma 2_1, \sigma 2_2$) are estimated with less bias than the level-3 standard deviations ($\sigma 3, \sigma 4_k$). Among the level-3 standard deviation parameters, the relative bias of $\sigma 3$ is smaller than that of $\sigma 4_k$.

Table 1. Statistics of Gibbs sampling of standard deviation parameters under all conditions.

Parameter	True value	$n = 50,\ s = 20$		$n = 20,\ s = 50$		$n = 40,\ s = 40$	
		Mean	SD	Mean	SD	Mean	SD
$\sigma 2_1$	1.0	1.0251	0.0659	1.0138	0.0677	1.0085	0.0579
$\sigma 2_2$	1.0	1.0361	0.0698	1.0268	0.0717	0.9973	0.0543
$\sigma 3$	2.0	2.1733	0.3971	2.1005	0.2332	2.0881	0.2546
$\sigma 4_1$	0.2	0.3092	0.1897	0.3194	0.1937	0.2506	0.1482
$\sigma 4_2$	0.2	0.3201	0.1906	0.3192	0.1933	0.2665	0.1500
$\sigma 4_3$	1.0	1.1107	0.3170	1.0666	0.2602	1.0490	0.2111
$\sigma 4_4$	0.2	0.3223	0.2050	0.3197	0.2032	0.2680	0.1603
$\sigma 4_5$	1.0	1.1185	0.3164	1.0666	0.2457	1.0499	0.2074
$\sigma 4_6$	0.2	0.3259	0.1976	0.3361	0.2000	0.2605	0.1530
$\sigma 4_7$	0.2	0.3053	0.1916	0.3299	0.1956	0.2469	0.1499
$\sigma 4_8$	0.2	0.3139	0.1944	0.3579	0.2049	0.2626	0.1538
$\sigma 4_9$	0.2	0.3350	0.2152	0.3298	0.2065	0.2797	0.1652
$\sigma 4_{10}$	0.2	0.3035	0.2048	0.3557	0.2130	0.2957	0.1688

Table 2. Estimates of the percentage relative bias for the standard deviation parameters under all conditions.

Parameter	True value	Relative bias %		
		$n = 50,\ s = 20$	$n = 20,\ s = 50$	$n = 40,\ s = 40$
$\sigma 2_1$	1.0	2.51	1.38	0.85
$\sigma 2_2$	1.0	3.61	2.68	-0.27
$\sigma 3$	2.0	8.67	5.03	4.41
$\sigma 4_1$	0.2	54.6	59.70	25.3
$\sigma 4_2$	0.2	60.05	59.60	33.25
$\sigma 4_3$	1.0	11.07	6.66	4.90
$\sigma 4_4$	0.2	61.15	59.85	34.00
$\sigma 4_5$	1.0	11.85	6.66	4.99
$\sigma 4_6$	0.2	62.95	68.05	30.25
$\sigma 4_7$	0.2	52.65	64.95	23.45
$\sigma 4_8$	0.2	56.95	78.95	31.30
$\sigma 4_9$	0.2	67.50	64.90	39.85
$\sigma 4_{10}$	0.2	51.75	77.85	47.55

"Relative bias %" of σ is computed using $[(\overline{\hat{\sigma}} - \sigma)/\sigma] \times 100\%$

4. Conclusion and Suggestion

From these results, we can conclude that the estimates for the standard deviations are positively biased. The relative bias of standard deviation estimates is inversely related to amount of information in the data that they are based on, and their magnitudes. The positive bias of the standard deviation parameters can be explained from the Bayesian point of view. Bayesian estimation combines the prior distribution with the likelihood to obtain the posterior distribution. When the data provide less information, the posterior distribution more heavily weights the prior resulting in shrinkage of the standard deviations toward the mean of the prior, which is a large positive value for a uniform distribution with a wide range from 0 to 1000 or a half-Cauchy distribution.

The bias can be reduced by increasing the number of schools and the number of students per school as can be seen from Table 2. In addition, the skewness of the posterior density estimate for $\sigma 4_1$ shown in Figure 3 suggests that the use of the median or mode of its posterior density as a point estimate may also reduce the relative bias.

References

[1] Bafumi, J., Gelman, A., Park, D.K., & Kaplan N. Practical issues in implementing and understanding bayesian ideal point estimation. *Political Analysis*, **13**; 2005: 171-187.

[2] Bolt, D. A SIBTEST approach to testing DIF hypotheses using experimentally designed test items. *Journal of Educational Measurement*, **37**; 2000: 307-327.

[3] Gelman, A. Prior distributions for variance parameters in hierarchical models. Submitted. Downloadable from **http://www.stat.columbia.edu/~gelman/research/published/tau9.pdf**, [2004, April 10].

[4] Gierl, M.J., Bisanz, J., Bisanz, G.L., & Boughton, K.A. Identifying content and cognitive skills that procedure gender differences in mathematics: A demonstration of the multidimensionality-based DIF analysis paradigm. *Journal of Educational Measurement*}, **4**; 2003: 281-306.

[5] Kamata, A., & Binici, S. Random-effect DIF analysis via hierarchical generalized linear models. Paper presented at the Annual International Meeting of the Psychometric Society, Sardinia, Italy. 2003.

[6] Millsap, R.E., Everson, H.T. Methodology review: Statistical approaches for assessing measurement bias. *Applied Psychological Measurement* **17**; 1993: 297-334.

[7] Swaminathan, H., & Rogers, H.J. Detecting differential item functioning using logistic regression procedures. *Journal of Educational Measurement*, **27**; 1990: 361-370.

[8] Swanson, D.B., Clauser, B.E., Case, S.M., Nungester, R.J., & Featherman, C. Analysis of differential item functioning (DIF) using hierarchical logistic regression models. *Journal of Educational and Behavioral Statistics*, **27**; 2002: 53-75.